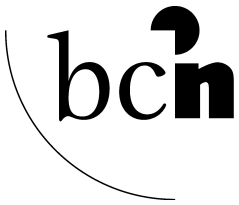
Data-driven identification of fixed expressions and their modifiability

Villada Moirón, María Begoña

Data-driven identification of fixed expressions and their modifiability

María Begoña Villada Moirón



The work in this thesis has been carried out under the auspices of the Behavioral and Cognitive Neurosciences (BCN) research school, Groningen, and has been part of the PIONIER project *Algorithms for Linguistic Processing* supported by grant number 220-70-001 from the Netherlands Organization for Scientific Research (NWO).

Groningen Dissertations in Linguistics 52
ISSN 0928-0030
Document prepared with LaTeX 2_ε
Cover designed by Misha Scholten. Printed by PrintPartners Ipskamp.

Rijks*universiteit* Groningen

Data-driven identification of fixed expressions and their modifiability

Proefschrift

ter verkrijging van het doctoraat in de
Letteren
aan de Rijksuniversiteit Groningen
op gezag van de
Rector Magnificus, dr. F. Zwarts,
in het openbaar te verdedigen op
donderdag 24 maart 2005
om 13.15 uur

door

María Begoña Villada Moirón

geboren op 11 juli 1972
te Riotorto, Spain

Promotor:	Prof. dr. ir. J. Nerbonne
Copromotores:	Dr. G. van Noord Dr. G. Bouma
Beoordelingscommissie:	Prof. J. Hoeksema Prof. J. Odijk Prof. I. Sag

Preface

During the years that I spent as a PhD student at the Alfa-Informatica department I met many people to whom I am truly grateful and who had a major impact on my thesis, my life, or both.

This thesis would never have come into being without the invaluable advice, guidance and support of my two supervisors: Gosse Bouma and Gertjan van Noord. They always found time to listen to the many difficulties I encountered along the way and discuss solutions. From countless discussions with them, I learnt a lot about scientific research, statistics, computational linguistics and (aghhh, I even have to mention them in the preface!) fixed expressions. Thank you both for your crucial input and the many corrections of former drafts. And, thank you for letting me work at my own pace and pushing me when necessary.

I owe a heavy debt to my promotor John Nerbonne whose wisdom and suggestions jotted down in numerous drafts I tried to adopt, although not always succeeding in covering them all. Throughout the whole project, John helped me with the statistics; he also read drafts amazingly quickly and returned insightful feedback, as well as corrected my English writing. Of course, thank you for providing extra financial support when needed and for maintaining a healthy research environment.

To the list of colleagues deserving of gratitude, I'd like to add Leonoor van der Beek who always agreed to read my writings and returned criticism and interesting suggestions for improvement. Leonoor always found a place in her agenda to discuss my work and the so much needed linguistic data. I also owe her for the Dutch translation of the thesis summary. Thanks also to Lonneke van der Plas for her contribution.

Being a non-native and clumsy Dutch speaker posed difficulties to evaluate the output of my experiments. Without hesitation, Gosse Bouma, Gertjan van Noord, and Leonoor van der Beek, stoically assessed long lists of (sometimes nonsensical) expressions and always gave me constructive feedback. I'm also grateful to Robbert Prins, Henny Klein, Jack Hoeksema, Bart Hollebrandse and Dirk-Bart den Ouden for assessing and discussing the

specific language data.

I am also indebted to the members of my reading committee: Jack Hoeksema, Jan Odijk and Ivan Sag. They corrected various mistakes in a previous manuscript and made detailed and helpful suggestions for improvement that (to some extent) were incorporated in this book. None of the above is of course responsible for how the thesis turned out in the end; the remaining mistakes or inaccuracies are entirely my own responsibility.

Many other people also helped me in many ways. I would like to acknowledge the members of the PIONIER project *Algorithms for Linguistic Processing* for all their instruction and discussions, in particular its former members Tanja Gaustad, Jan Daciuk and Rob Malouf. My work also benefited from the input of many colleagues who participated in various reading groups and contributed to many fruitful exchanges. In addition, I am grateful to my (through-various-years) office mates: Rob Malouf, Stasinos Konstantopoulos, Susanne Schoof, Holger Hopp and Jörg Tiedemann for their assistance with a variety of technical matters, thought-provoking discussions, or simply, their occasional interruptions to remind me to take a coffee break. Stas deserves separate mention for creating the RuG thesis style file. A big thanks to Mark-Jan Nederhoff for sharing his substantial knowledge of LaTeX and book typesetting. I also would like to acknowledge the encouragement and support of my colleagues in Alfa-Informatica who make the department such a stimulating and enjoyable environment for research. Last but not least, I thank the fourth floor secretarial staff for their assistance with the day-to-day issues and our system administrators for their computer support.

Among those less directly related to this work, I am grateful to Carl Vogel who encouraged me to continue doing research in computational linguistics, even if that meant abandoning the at-the-time 'dreadful' *clitics*. Also, thanks to the Hungarian members of the OTKA-NWO project *Finding and Processing Multi-word Lexemes* for testing my methods on a totally different language, with different settings and application, as well as for many hours involved in e-discussions.

Extra special thanks to Leonoor van der Beek and Francisco Borges who kindly agreed to be my 'paranimfs'.

Como sempre, agradezo inmensamente a axuda da miña familia e é unha honra que poidan asistir á miña defensa de doutoramento. Finally I owe my greatest debt to Michiel for his help, support and willingness to be distracted from his own studies and work by talk of statistics and linguistics.

Contents